

# Levering Bulk Data to Build a Single-Cell Methylation Clock

Lisa Barooah<sup>1</sup>, Jason Ernst<sup>2</sup>

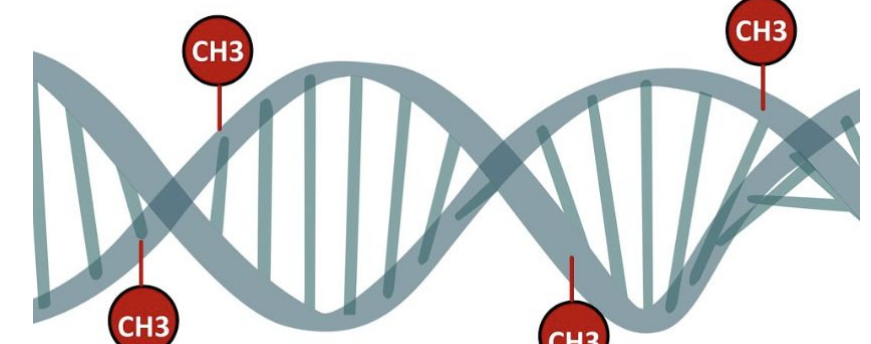
<sup>1</sup>BIG Summer Program, UCLA <sup>2</sup>Department of Computational Medicine, David Geffen School of Medicine, UCLA

## Abstract

DNA methylation clocks in aging research are traditionally built from bulk tissue. For instance, the widely used **Horvath clock** integrates information from 353 CpG sites to make predictions highly correlated with chronological age. Recently, single-cell methylation clocks have emerged but face coverage limitations during direct application to bulk data. In this project, we designed an approach to apply clocks trained on bulk data for application to pseudobulked single-cell datasets. To **address coverage issues**, we cross-referenced Horvath's 353 CpG sites with a compendium of bulk methylation profiles. We identified similar CpG sites by calculating the smallest Euclidean distances and used weighted averages of their values as features for prediction. This approach mitigates imputation challenges inherent to single-cell data while enabling more accurate age estimation.

## Introduction

DNA methylation is a robust biomarker of biological aging and forms the basis of epigenetic clocks. Fig 1. Animated Biology with Arpan, DNA Methylation, 2022



In this study, we adapt the Horvath Clock to single-cell bisulfite sequencing data. Although clocks like scAge and scEpiAge have been published recently, single-cell methylation clocks still struggle with coverage issues and reproducibility. We aim to use bulk data to provide a more reproducible single-cell epigenetic clock.

## Materials

- Datasets:
- Clustered Bisulfite Data from Single-Cell Hg38 Data
    - 4 sample individuals (C29, C37, C38, C39)
    - Adult donors (aged 44-55 years)
    - Fibroblasts that underwent cellular reprogramming
  - Predicted Clustering Order:
    - early
    - sendai
    - skin
    - mid1
    - inflamed
    - fail1
    - fail2
    - mid2
    - ips
  - Horvath Clock CpG Site Data in Hg19
  - Bulk WGBS Methylation Data from IHEC
    - 646 sample individuals for each CpG site
    - Accessed via International Human Epigenome Consortium

## Methodology

- Part 1: Data Standardization**
- Converted Horvath hg18 mapping to hg38 mapping using UCSC Liftover
- Part 2: Parse through all samples from pseudobulked single cell data**
- Found Horvath CpG sites and their methylation data
- Part 3: Calculate predicted biological ages**
- Found horvath scores, then converted to predicted biological age using a linear transformation
- Referred to this as the trivial method

--- Match at chr1:1232656 ---					
chr1	1232645	+	CCC	0	28
chr1	1232646	+	CCG	0	29
chr1	1232647	+	CGC	11	29
chr1	1232648	-	CGG	4	26
chr1	1232649	+	CGC	12	28
chr1	1232650	-	CGC	8	27
chr1	1232651	+	CTG	0	26
chr1	1232652	-	CAG	0	26
chr1	1232653	-	CAG	0	26
chr1	1232654	+	CGC	7	26
chr1	1232655	-	CGC	5	28
chr1	1232656	+	CGA	12	26
chr1	1232657	-	CGC	5	28
chr1	1232658	-	CGC	10	28
chr1	1232659	+	CGC	10	28

## -Part 4: Develop window and strand-aware imputation strategies in an attempt to remedy low coverage

Window					
chr #	base pair #	strand	sequence	methylation reads	total reads
--- Match at chr1:1232656 ---					
chr1	1232645	+	CCC	0	28
chr1	1232646	+	CCG	0	29
chr1	1232647	+	CGC	11	29
chr1	1232648	-	CGG	4	26
chr1	1232649	+	CGC	12	28
chr1	1232650	-	CGC	8	27
chr1	1232651	+	CTG	0	26
chr1	1232652	-	CAG	0	26
chr1	1232653	-	CAG	0	26
chr1	1232654	+	CGC	7	26
chr1	1232655	-	CGC	5	28
chr1	1232656	+	CGA	12	26
chr1	1232657	-	CGC	5	28
chr1	1232658	-	CGC	10	28
chr1	1232659	+	CGC	10	28

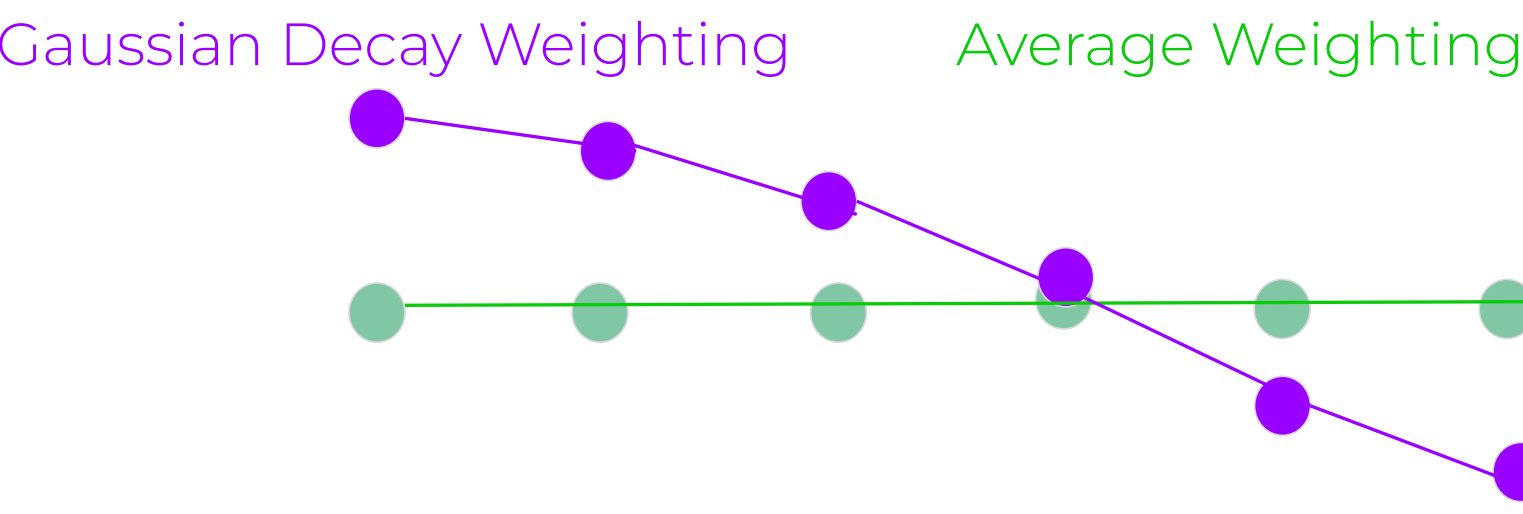
Strand Aware					
chr #	base pair #	strand	sequence	methylation reads	total reads
--- Match at chr1:1232656 ---					
chr1	1232645	+	CCC	0	28
chr1	1232646	+	CCG	0	29
chr1	1232647	+	CGC	11	29
chr1	1232648	-	CGG	4	26
chr1	1232649	+	CGC	12	28
chr1	1232650	-	CGC	8	27
chr1	1232651	+	CTG	0	26
chr1	1232652	-	CAG	0	26
chr1	1232653	-	CAG	0	26
chr1	1232654	+	CGC	7	26
chr1	1232655	-	CGC	5	28
chr1	1232656	+	CGA	12	26
chr1	1232657	-	CGC	5	28
chr1	1232658	-	CGC	10	28
chr1	1232659	+	CGC	10	28

## Part 5: Identify Top Similar CpG Sites from IHEC Data

Identified top similar CpG sites based on the smallest euclidean distance from Horvath profiles

## -Part 5: Applied the clock after smoothing CpG values

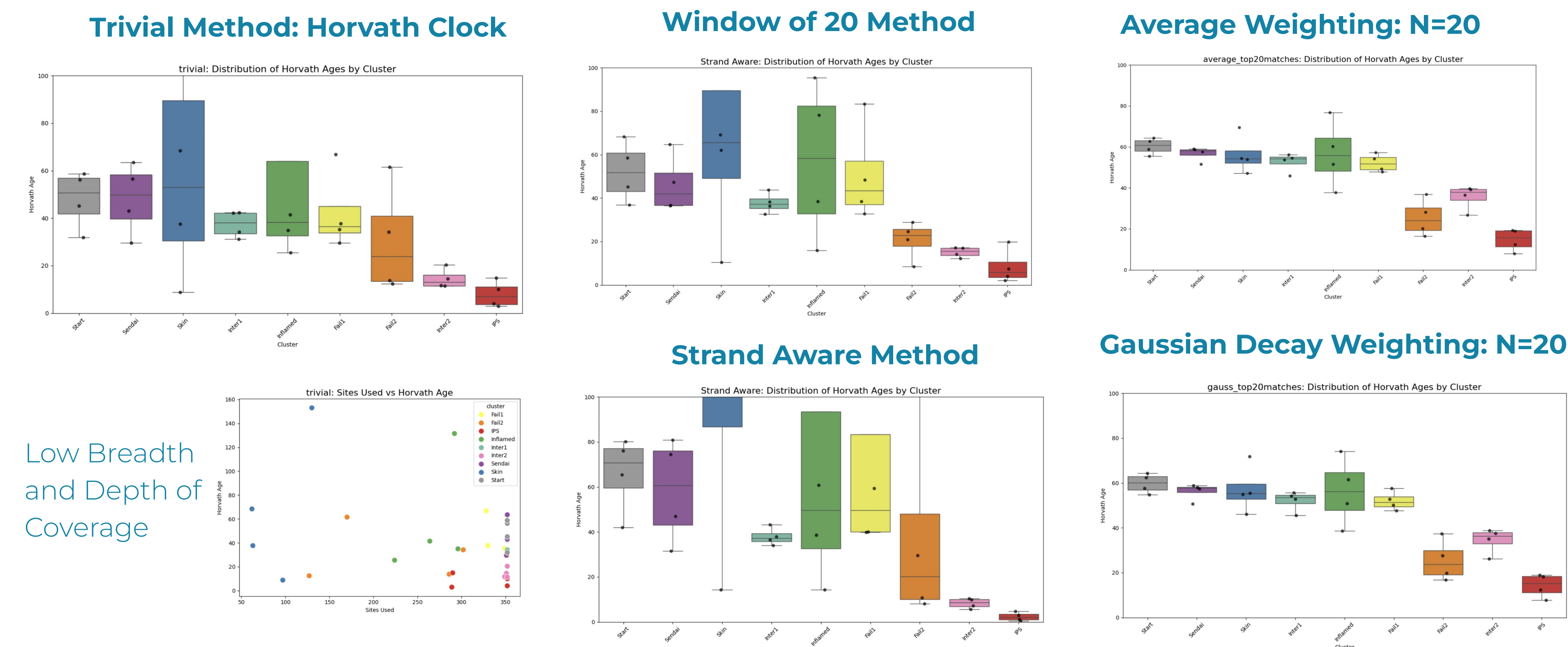
Applied two weighted smoothing strategies:



horvath site: match 1, match 2, match 3, match 4, match 5, match 6...

## Results

### Comparing Different Methods Biological Age Prediction



### Spearman Analysis of Methods

Method	Spearman_r	#	p_value	Avg_CpGs	Avg_IntraCluster_Stk
strandaware_summary	-0.5833	0.0992	351.97	42.9	
window_size20_summary	-0.7167	0.0298	299.53	18.68	
trivial_summary	-0.7667	0.0159	293.94	21.55	
average_top20matches_summary	-0.9333	0.0002	351.97	6.98	
gauss_top20matches_summary	-0.95	0.0001	351.97	7	

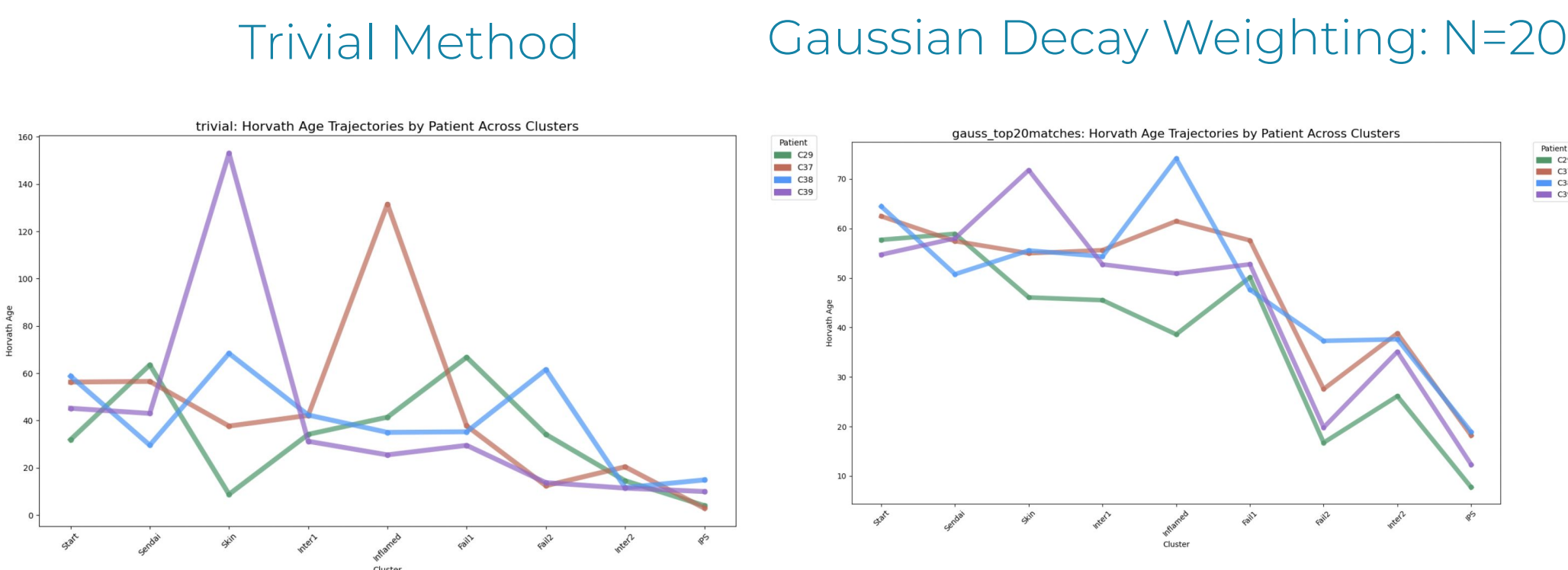
Imputing with similar CpG sites improved the correlation among clusters by from **-0.76 to -0.95**. Using similar CpG sites improves coverage issues and more effectively differentiates clusters.

The negative correlation is expected due to the nature of our project's clustering. The clock has a significant p-value and lower intra-cluster standard deviation.

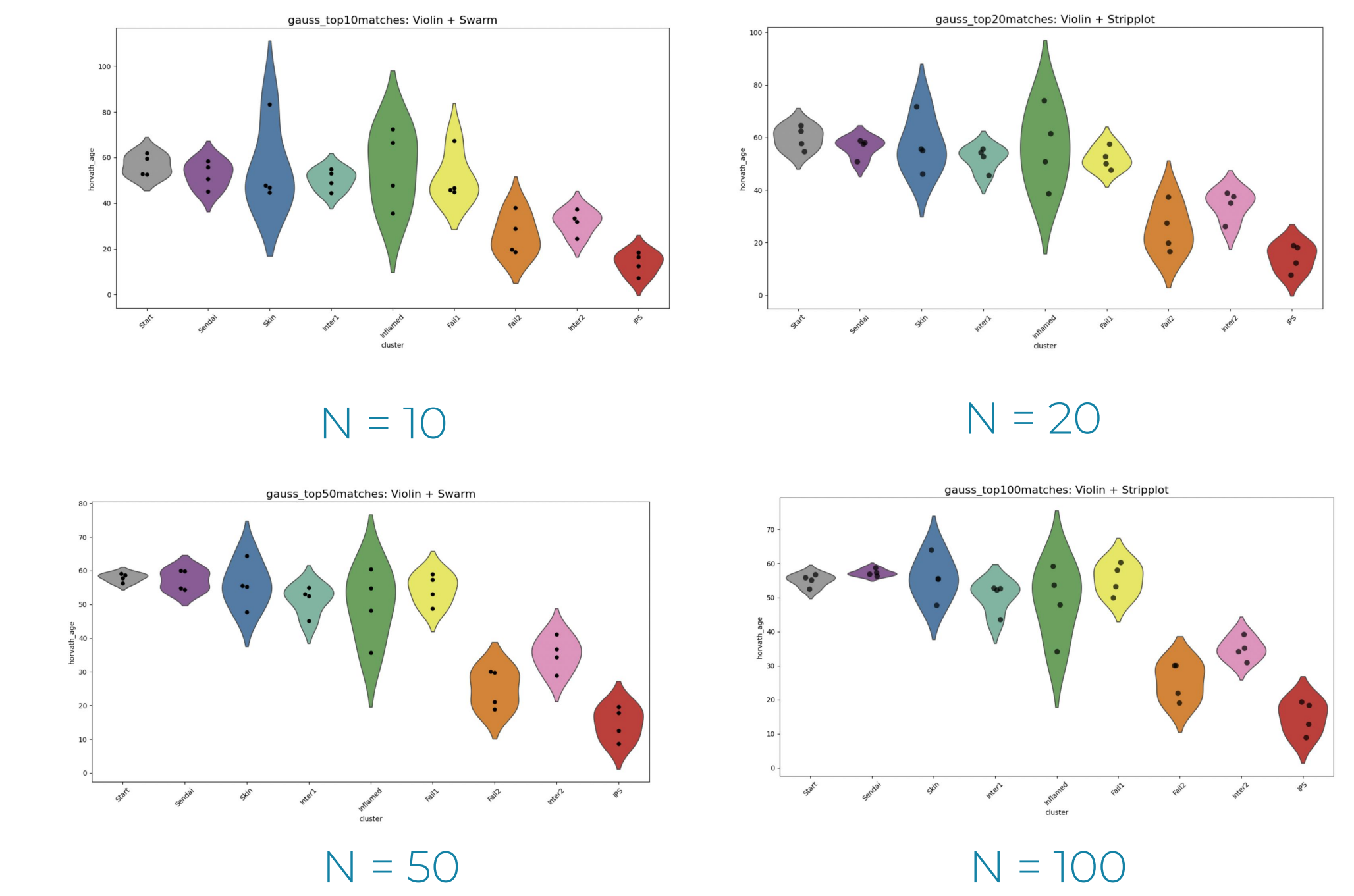
Method	Spearman_r	#	p_value	Avg_CpGs	Avg_IntraCluster_Stk
trivial_summary	-0.7667	0.0159	293.94	21.55	
gauss_strandaware_top100matche	-0.7833	0.0125	351.97	9.15	
gauss_window20_top100matches	-0.8	0.0096	352.72	4.84	
gauss_top100matches	-0.8	0.0096	351.97	4.83	
gauss_window20_top20matches	-0.8	0.0096	352.72	4.84	
gauss_window20_top10matches	-0.8	0.0096	352.72	4.84	
gauss_strandaware_top20match	-0.8333	0.0053	351.97	13.56	
gauss_strandaware_top10matches	-0.8333	0.0053	351.97	18.76	
gauss_top10matches	-0.8667	0.0025	351.97	8.97	
gauss_top50matches	-0.9333	0.0002	351.97	5.17	
gauss_top40matches	-0.9333	0.0002	351.97	5.2	
gauss_top20matches	-0.95	0.0001	351.97	7	

## Results

### Order Preservation in Methods



### Comparison of Different Top Match Sizes



## Conclusion

Our analysis suggests that **using sites that have been proven similar to the previously identified Horvath sites can improve predictions**, as these regions may provide more reliable methylation signals. Both the gaussian and average weightings dramatically improved correlation values.

In contrast, window-based methods and the “+1” approach actually worsened performance, likely because single-cell methylation data are inherently sparse.

We hope to test across additional donors and timepoints to determine how early in reprogramming biological age reduction begins.

## Acknowledgements

Thank you to Professor Jason Ernst and the Computational Medicine team at the UCLA Geffen School of Medicine for assisting me in this project! The bulk WGBS data was compiled by the International Human Epigenome Consortium (IHEC). The IGVF single cell data was compiled by the UCLA Impact of Genome Variation on Function Group. Special thanks to Dr. Hoffman and the B.I.G. Summer team as well!